Econ 423 – Lecture Notes

(These notes are slightly modified versions of lecture notes provided by Stock and Watson, 2007. They are for instructional purposes only and are not to be distributed outside of the classroom.)

Instrumental Variables Regression

Three important threats to internal validity are:

- omitted variable bias from a variable that is correlated with X but is unobserved, so cannot be included in the regression;
- simultaneous causality bias (X causes Y, Y causes X);
- errors-in-variables bias (X is measured with error)

Instrumental variables regression can eliminate bias when $E(u|X) \neq 0$ – using an *instrumental variable*, Z

IV Regression with One Regressor and One Instrument

$$Y_i = \beta_0 + \beta_1 X_i + u_i$$

- IV regression breaks X into two parts: a part that might be correlated with u, and a part that is not. By isolating the part that is not correlated with u, it is possible to estimate β_1.
- This is done using an ***instrumental variable***, Z_i, which is uncorrelated with u_i.
- The instrumental variable detects movements in X_i that are uncorrelated with u_i, and uses these to estimate β_1.

Terminology: endogeneity and exogeneity

An ***endogenous*** variable is one that is correlated with *u*
An ***exogenous*** variable is one that is uncorrelated with *u*

> *Historical note:* "Endogenous" literally means "determined within the system," that is, a variable that is jointly determined with Y, that is, a variable subject to simultaneous causality. However, this definition is narrow and IV regression can be used to address OV bias and errors-in-variable bias, not just to simultaneous causality bias.

Two conditions for a valid instrument

$$Y_i = \beta_0 + \beta_1 X_i + u_i$$

For an instrumental variable (an "***instrument***") Z to be valid, it must satisfy two conditions:

1. ***Instrument relevance***: $\text{corr}(Z_i,X_i) \neq 0$
2. ***Instrument exogeneity***: $\text{corr}(Z_i,u_i) = 0$

Suppose for now that you have such a Z_i (we'll discuss how to find instrumental variables later).

How can you use Z_i to estimate β_1?

The IV Estimator, one *X* and one *Z*

Explanation #1: Two Stage Least Squares (TSLS)

As it sounds, TSLS has two stages – two regressions:

(1) First isolates the part of *X* that is uncorrelated with *u*: regress *X* on *Z* using OLS

$$X_i = \pi_0 + \pi_1 Z_i + v_i \tag{1}$$

- Because Z_i is uncorrelated with u_i, $\pi_0 + \pi_1 Z_i$ is uncorrelated with u_i. We don't know π_0 or π_1 but we have estimated them, so…
- Compute the predicted values of X_i, $\hat{X}_i$, where $\hat{X}_i = \hat{\pi}_0 + \hat{\pi}_1 Z_i$, $i = 1,\ldots,n$.

Two Stage Least Squares, ctd.

(2) Replace X_i by $\hat{X}_i$ in the regression of interest:

regress Y on $\hat{X}_i$ using OLS:

$$Y_i = \beta_0 + \beta_1 \hat{X}_i + u_i \qquad (2)$$

- **Because $\hat{X}_i$ is uncorrelated with u_i (if n is large), the first least squares assumption holds (if n is large)**
- Thus β_1 can be estimated by OLS using regression (2)
- This argument relies on large samples (so π_0 and π_1 are well estimated using regression (1))
- This the resulting estimator is called the *Two Stage Least Squares (TSLS)* estimator, $\hat{\beta}_1^{TSLS}$.

Two Stage Least Squares, ctd.

Suppose you have a valid instrument, Z_i.

Stage 1: Regress X_i on Z_i, obtain the predicted values $\hat{X}_i$

Stage 2: Regress Y_i on $\hat{X}_i$; the coefficient on $\hat{X}_i$ is the TSLS estimator, $\hat{\beta}_1^{TSLS}$.

$\hat{\beta}_1^{TSLS}$ is a consistent estimator of β_1.

The IV Estimator, one X and one Z, ctd.

Explanation #2: a little algebra…

$$Y_i = \beta_0 + \beta_1 X_i + u_i$$

Thus,

$$\begin{aligned}
\text{cov}(Y_i,Z_i) &= \text{cov}(\beta_0 + \beta_1 X_i + u_i, Z_i) \\
&= \text{cov}(\beta_0,Z_i) + \text{cov}(\beta_1 X_i,Z_i) + \text{cov}(u_i,Z_i) \\
&= \quad 0 \quad + \text{cov}(\beta_1 X_i,Z_i) + \quad 0 \\
&= \beta_1 \text{cov}(X_i,Z_i)
\end{aligned}$$

where $\text{cov}(u_i,Z_i) = 0$ (instrument exogeneity); thus

$$\beta_1 = \frac{\text{cov}(Y_i, Z_i)}{\text{cov}(X_i, Z_i)}$$

The IV Estimator, one X and one Z, ctd.

$$\beta_1 = \frac{\text{cov}(Y_i, Z_i)}{\text{cov}(X_i, Z_i)}$$

The IV estimator replaces these population covariances with sample covariances:

$$\hat{\beta}_1^{TSLS} = \frac{s_{YZ}}{s_{XZ}},$$

s_{YZ} and s_{XZ} are the sample covariances. This is the TSLS estimator – just a different derivation!

Consistency of the TSLS estimator

$$\hat{\beta}_1^{TSLS} = \frac{s_{YZ}}{s_{XZ}}$$

The sample covariances are consistent: $s_{YZ} \xrightarrow{p} \text{cov}(Y,Z)$ and $s_{XZ} \xrightarrow{p} \text{cov}(X,Z)$. Thus,

$$\hat{\beta}_1^{TSLS} = \frac{s_{YZ}}{s_{XZ}} \xrightarrow{p} \frac{\text{cov}(Y,Z)}{\text{cov}(X,Z)} = \beta_1$$

- The instrument relevance condition, $\text{cov}(X,Z) \neq 0$, ensures that you don’t divide by zero.

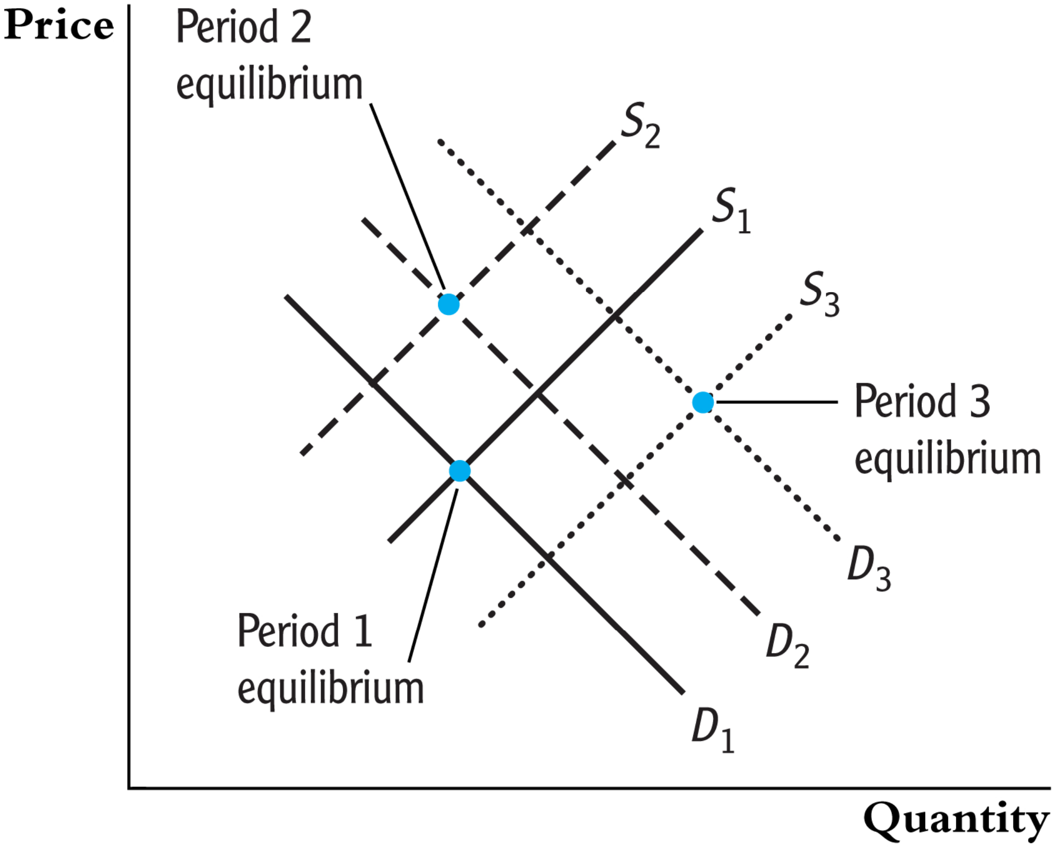
Example #1: Supply and demand for butter

IV regression was originally developed to estimate demand elasticities for agricultural goods, for example butter:

$$\ln(Q_i^{butter}) = \beta_0 + \beta_1 \ln(P_i^{butter}) + u_i$$

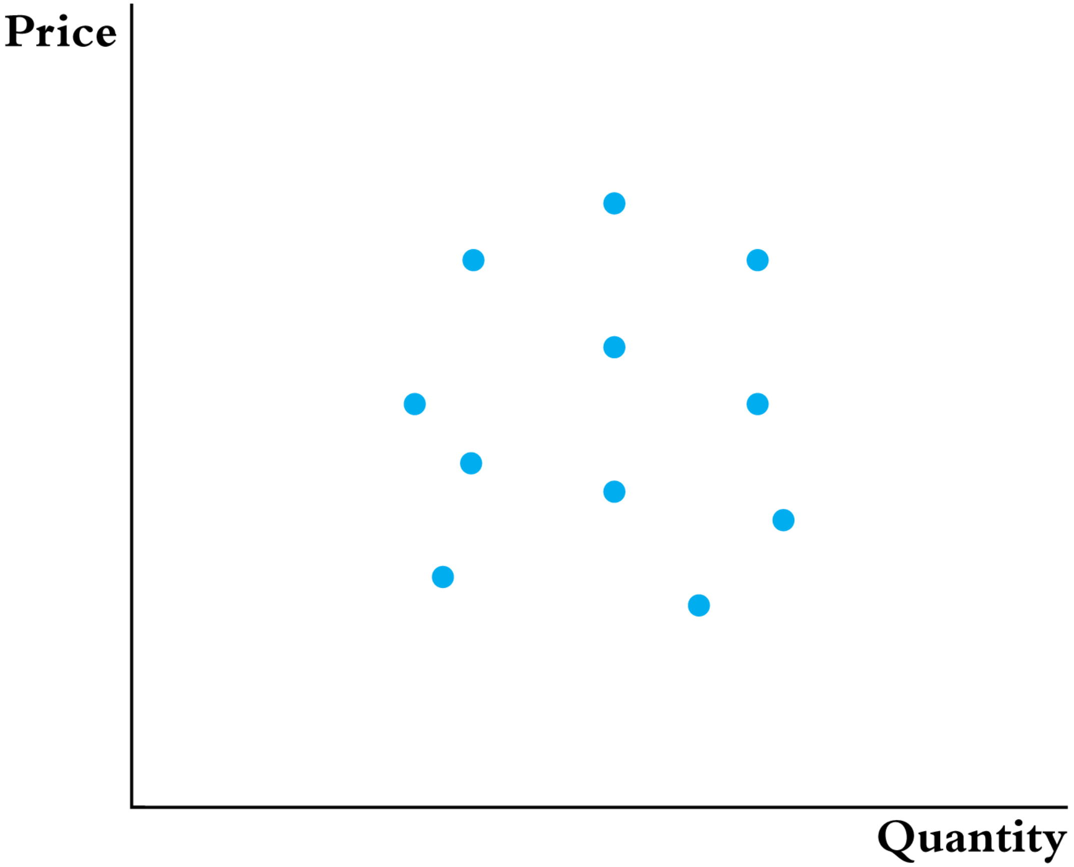
- β_1 = price elasticity of butter = percent change in quantity for a 1% change in price (recall log-log specification discussion)
- Data: observations on price and quantity of butter for different years
- The OLS regression of $\ln(Q_i^{butter})$ on $\ln(P_i^{butter})$ suffers from simultaneous causality bias (*why*?)

Simultaneous causality bias in the OLS regression of $\ln(Q_i^{butter})$ on $\ln(P_i^{butter})$ arises because price and quantity are determined by the interaction of demand *and* supply



(a) Demand and supply in three time periods

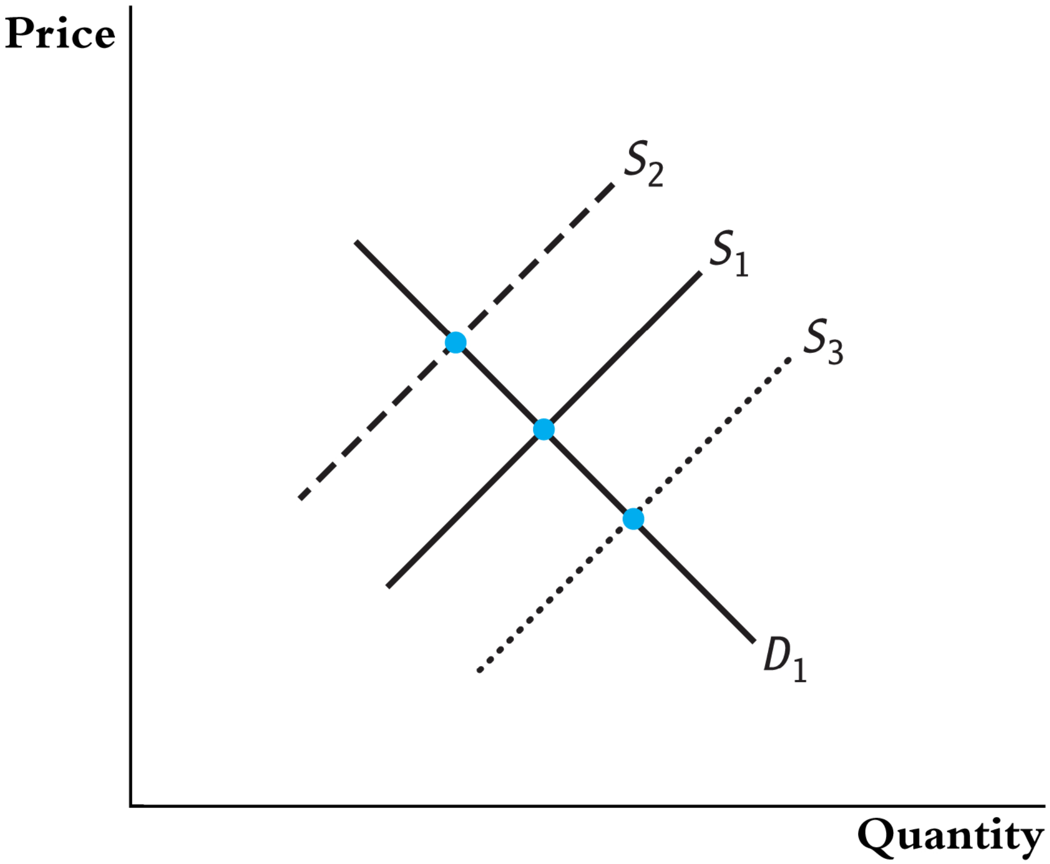
This interaction of demand and supply produces…



(b) Equilibrium price and quantity for 11 time periods

Would a regression using these data produce the demand curve?

But…what would you get if only supply shifted?



(c) Equilibrium price and quantity when only the supply curve shifts

- TSLS estimates the demand curve by isolating shifts in price and quantity that arise from shifts in supply.
- *Z* is a variable that shifts supply but not demand.

TSLS in the supply-demand example:

$$\ln(Q_i^{butter}) = \beta_0 + \beta_1 \ln(P_i^{butter}) + u_i$$

Let Z = rainfall in dairy-producing regions.

Is Z a valid instrument?

(1) Exogenous? corr($rain_i$,u_i) = 0?

Plausibly: whether it rains in dairy-producing regions shouldn’t affect demand

(2) Relevant? corr($rain_i$,ln(P_i^{butter})) ≠ 0?

Plausibly: insufficient rainfall means less grazing means less butter

TSLS in the supply-demand example, ctd.

$$\ln(Q_i^{butter}) = \beta_0 + \beta_1 \ln(P_i^{butter}) + u_i$$

$Z_i = rain_i$ = rainfall in dairy-producing regions.

Stage 1: regress $\ln(P_i^{butter})$ on *rain*, get $\widehat{\ln}(P_i^{butter})$

$\widehat{\ln}(P_i^{butter})$ isolates changes in log price that arise from supply (part of supply, at least)

Stage 2: regress $\ln(Q_i^{butter})$ on $\widehat{\ln}(P_i^{butter})$

The regression counterpart of using shifts in the supply curve to trace out the demand curve.

Example #2: Test scores and class size

- The California regressions still could have OV bias (e.g. parental involvement).
- This bias could be eliminated by using IV regression (TSLS).
- IV regression requires a valid instrument, that is, an instrument that is:
  (1) relevant: corr(Z_i,STR_i) $\neq 0$
  (2) exogenous: corr(Z_i,u_i) $= 0$

Example #2: Test scores and class size, ctd.

Here is a (hypothetical) instrument:

- some districts, randomly hit by an earthquake, "double up" classrooms:

$$Z_i = Quake_i = 1 \text{ if hit by quake, } = 0 \text{ otherwise}$$

- *Do the two conditions for a valid instrument hold*?
- The earthquake makes it *as if* the districts were in a random assignment experiment. Thus the variation in *STR* arising from the earthquake is exogenous.
- The first stage of TSLS regresses *STR* against *Quake*, thereby isolating the part of *STR* that is exogenous (the part that is "as if" randomly assigned)

Inference using TSLS

- In large samples, the sampling distribution of the TSLS estimator is normal
- Inference (hypothesis tests, confidence intervals) proceeds in the usual way, e.g. ± $1.96SE$
- The idea behind the large-sample normal distribution of the TSLS estimator is that – like all the other estimators we have considered – it involves an average of mean zero i.i.d. random variables, to which we can apply the CLT.
- Here is a sketch of the math (see SW App. 12.3 for the details)...

$$\hat{\beta}_1^{TSLS} = \frac{s_{YZ}}{s_{XZ}} = \frac{\frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})(Z_i - \bar{Z})}{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})(Z_i - \bar{Z})}$$

$$= \frac{\sum_{i=1}^{n} Y_i(Z_i - \bar{Z})}{\sum_{i=1}^{n} X_i(Z_i - \bar{Z})}$$

Substitute in $Y_i = \beta_0 + \beta_1 X_i + u_i$ and simplify:

$$\hat{\beta}_1^{TSLS} = \frac{\beta_1\sum_{i=1}^{n} X_i(Z_i - \bar{Z}) + \sum_{i=1}^{n} u_i(Z_i - \bar{Z})}{\sum_{i=1}^{n} X_i(Z_i - \bar{Z})}$$

so…

$$\hat{\beta}_1^{TSLS} = \beta_1 + \frac{\sum_{i=1}^{n} u_i (Z_i - \bar{Z})}{\sum_{i=1}^{n} X_i (Z_i - \bar{Z})}.$$

so $$\hat{\beta}_1^{TSLS} - \beta_1 = \frac{\sum_{i=1}^{n} u_i (Z_i - \bar{Z})}{\sum_{i=1}^{n} X_i (Z_i - \bar{Z})}$$

Multiply through by $\sqrt{n}$:

$$\sqrt{n}(\hat{\beta}_1^{TSLS} - \beta_1) = \frac{\frac{1}{\sqrt{n}}\sum_{i=1}^{n} (Z_i - \bar{Z}) u_i}{\frac{1}{n}\sum_{i=1}^{n} X_i (Z_i - \bar{Z})}$$

$$\sqrt{n}\,(\hat{\beta}_1^{TSLS} - \beta_1) = \frac{\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(Z_i - \bar{Z})u_i}{\frac{1}{n}\sum_{i=1}^{n}X_i(Z_i - \bar{Z})}$$

- $\frac{1}{n}\sum_{i=1}^{n}X_i(Z_i - \bar{Z}) = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})(Z_i - \bar{Z}) \xrightarrow{p} \text{cov}(X,Z) \neq 0$
- $\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(Z_i - \bar{Z})u_i$ is dist'd $N(0,\text{var}[(Z–\mu_Z)u])$ (CLT)

so: $\hat{\beta}_1^{TSLS}$ is approx. distributed $N(\beta_1, \sigma^2_{\hat{\beta}_1^{TSLS}})$,

where $\sigma^2_{\hat{\beta}_1^{TSLS}} = \frac{1}{n}\frac{\text{var}[(Z_i - \mu_Z)u_i]}{[\text{cov}(Z_i, X_i)]^2}$.

where cov(X,Z) ≠ 0 because the instrument is relevant

Inference using TSLS, ctd.

$$\hat{\beta}_1^{TSLS} \text{ is approx. distributed } N(\beta_1, \sigma^2_{\hat{\beta}_1^{TSLS}}),$$

- Statistical inference proceeds in the usual way.
- The justification is (as usual) based on large samples
- This all assumes that the instruments are valid – we'll discuss what happens if they aren't valid shortly.
- ***Important note on standard errors***:
  - The OLS standard errors from the second stage regression aren't right – they don't take into account the estimation in the first stage ($\hat{X}_i$ is estimated).
  - Instead, use a single specialized command that computes the TSLS estimator and the correct *SE*s.
  - as usual, use heteroskedasticity-robust *SE*s

Example: Cigarette demand, ctd.

$$\ln(Q_i^{cigarettes}) = \beta_0 + \beta_1 \ln(P_i^{cigarettes}) + u_i$$

Panel data:

- Annual cigarette consumption and average prices paid (including tax)
- 48 continental US states, 1985-1995

Proposed instrumental variable:

- Z_i = general sales tax per pack in the state = $SalesTax_i$
- Is this a valid instrument?

  (1) Relevant? corr($SalesTax_i$, $\ln(P_i^{cigarettes})$) $\neq 0$?

  (2) Exogenous? corr($SalesTax_i$,u_i) = 0?

Cigarette demand, ctd.

For now, use data from 1995 only.

First stage OLS regression:

$$\widehat{\ln(P_i^{cigarettes})} = 4.63 + .031 SalesTax_i,\ n = 48$$

Second stage OLS regression:

$$\widehat{\ln(Q_i^{cigarettes})} = 9.72 - 1.08\ \widehat{\ln(P_i^{cigarettes})},\ n = 48$$

Combined regression with correct, heteroskedasticity-robust standard errors:

$$\widehat{\ln(Q_i^{cigarettes})} = \underset{(1.53)}{9.72} - \underset{(0.32)}{1.08}\ \widehat{\ln(P_i^{cigarettes})},\ n = 48$$

STATA Example: Cigarette demand, First stage

Instrument = Z = *rtaxso* = general sales tax (real $/pack)

```
          X        Z
. reg lravgprs rtaxso if year==1995, r;

Regression with robust standard errors                 Number of obs =      48
                                                       F(  1,    46) =   40.39
                                                       Prob > F      =  0.0000
                                                       R-squared     =  0.4710
                                                       Root MSE      =  .09394

------------------------------------------------------------------------------
             |               Robust
    lravgprs |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      rtaxso |   .0307289   .0048354     6.35   0.000     .0209956    .0404621
       _cons |   4.616546   .0289177   159.64   0.000     4.558338    4.674755
------------------------------------------------------------------------------

             X-hat
. predict lravphat;          Now we have the predicted values from the 1st stage
```

Second stage

```
         Y       X-hat
. reg lpackpc lravphat if year==1995, r;

Regression with robust standard errors                 Number of obs =      48
                                                       F(  1,    46) =   10.54
                                                       Prob > F      =  0.0022
                                                       R-squared     =  0.1525
                                                       Root MSE      =  .22645

------------------------------------------------------------------------------
             |               Robust
     lpackpc |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    lravphat |  -1.083586   .3336949    -3.25   0.002    -1.755279   -.4118932
       _cons |   9.719875   1.597119     6.09   0.000     6.505042    12.93471
------------------------------------------------------------------------------
```

- These coefficients are the TSLS estimates
- The standard errors are wrong because they ignore the fact that the first stage was estimated

Combined into a single command:

```
           Y         X        Z
. ivreg lpackpc (lravgprs = rtaxso) if year==1995, r;

IV (2SLS) regression with robust standard errors       Number of obs =      48
                                                       F(  1,    46) =   11.54
                                                       Prob > F      =  0.0014
                                                       R-squared     =  0.4011
                                                       Root MSE      =  .19035

------------------------------------------------------------------------------
             |               Robust
     lpackpc |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    lravgprs |  -1.083587   .3189183    -3.40   0.001    -1.725536   -.4416373
       _cons |   9.719876   1.528322     6.36   0.000     6.643525    12.79623
------------------------------------------------------------------------------
Instrumented:  lravgprs          This is the endogenous regressor
Instruments:   rtaxso            This is the instrumental varible
------------------------------------------------------------------------------

       OK, the change in the SEs was small this time...but not always!
```

$$\widehat{\ln}(Q_i^{cigarettes}) = 9.72 - 1.08\, \widehat{\ln}(P_i^{cigarettes}),\ n = 48$$

$$(1.53)\quad (0.32)$$

Summary of IV Regression with a Single *X* and *Z*

- A valid instrument *Z* must satisfy two conditions:

  (1) *relevance*: corr(Z_i,X_i) ≠ 0

  (2) *exogeneity*: corr(Z_i,u_i) = 0
- TSLS proceeds by first regressing *X* on *Z* to get $\hat{X}$, then regressing *Y* on $\hat{X}$.
- The key idea is that the first stage isolates part of the variation in *X* that is uncorrelated with *u*
- If the instrument is valid, then the large-sample sampling distribution of the TSLS estimator is normal, so inference proceeds as usual

The General IV Regression Model

- So far we have considered IV regression with a single endogenous regressor (X) and a single instrument (Z).
- We need to extend this to:
  - multiple endogenous regressors ($X_1,\ldots,X_k$)
  - multiple included exogenous variables ($W_1,\ldots,W_r$)

    These need to be included for the usual OV reason
  - multiple instrumental variables ($Z_1,\ldots,Z_m$)

    More (relevant) instruments can produce a smaller variance of TSLS: the R^2 of the first stage increases, so you have more variation in $\hat{X}$.
- Terminology: identification & overidentification

Identification

- In general, a parameter is said to be ***identified*** if different values of the parameter would produce different distributions of the data.
- In IV regression, whether the coefficients are identified depends on the relation between the number of instruments (m) and the number of endogenous regressors (k)
- Intuitively, if there are fewer instruments than endogenous regressors, we can't estimate $\beta_1,\ldots,\beta_k$
  - For example, suppose $k = 1$ but $m = 0$ (no instruments)!

Identification, ctd.

The coefficients $\beta_1,\ldots,\ \beta_k$ are said to be:

- ***exactly identified*** if $m = k$.

  There are just enough instruments to estimate $\beta_1,\ldots,\beta_k$.

- ***overidentified*** if $m > k$.

  There are more than enough instruments to estimate $\beta_1,\ldots,\beta_k$. *If so, you can test whether the instruments are valid* (*a test of the "overidentifying restrictions*") *– we'll return to this later*

- ***underidentified*** if $m < k$.

  There are too few instruments to estimate $\beta_1,\ldots,\beta_k$. *If so, you need to get more instruments!*

The general IV regression model: Summary of jargon

$$Y_i = \beta_0 + \beta_1 X_{1i} + \ldots + \beta_k X_{ki} + \beta_{k+1} W_{1i} + \ldots + \beta_{k+r} W_{ri} + u_i$$

- Y_i is the ***dependent variable***
- $X_{1i},\ldots, X_{ki}$ are the ***endogenous regressors*** (potentially correlated with u_i)
- $W_{1i},\ldots,W_{ri}$ are the ***included exogenous variables*** or ***included exogenous regressors*** (uncorrelated with u_i)
- $\beta_0, \beta_1,\ldots, \beta_{k+r}$ are the unknown regression coefficients
- $Z_{1i},\ldots,Z_{mi}$ are the m ***instrumental variables*** (the ***excluded exogenous variables***)
- The coefficients are ***overidentified*** if $m > k$; ***exactly identified*** if $m =$ k; and ***underidentified*** if $m < k$.

TSLS with a single endogenous regressor

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 W_{1i} + \ldots + \beta_{1+r} W_{ri} + u_i$$

- m instruments: $Z_{1i},\ldots, Z_m$
- First stage
  - Regress X_1 on *all* the exogenous regressors: regress X_1 on $W_1,\ldots,W_r$, $Z_1,\ldots, Z_\mathrm{m}$ by OLS
  - Compute predicted values $\hat{X}_{1i}$, $i = 1,\ldots,n$
- Second stage
  - Regress Y on $\hat{X}_1$, $W_1,\ldots, W_r$ by OLS
  - The coefficients from this second stage regression are the TSLS estimators, but *SE*s are wrong
- To get correct *SE*s, do this in a single step

Example: Demand for cigarettes

$$\ln(Q_i^{cigarettes}) = \beta_0 + \beta_1\ln(P_i^{cigarettes}) + \beta_2\ln(Income_i) + u_i$$

$$Z_{1i} = \text{general sales tax}_i$$
$$Z_{2i} = \text{cigarette-specific tax}_i$$

- Endogenous variable: $\ln(P_i^{cigarettes})$ ("one X")
- Included exogenous variable: $\ln(Income_i)$ ("one W")
- Instruments (excluded endogenous variables): general sales tax, cigarette-specific tax ("two Zs")
- *Is the demand elasticity β_1 overidentified, exactly identified, or underidentified?*

Example: Cigarette demand, one instrument

```
                Y        W         X          Z
. ivreg lpackpc lperinc (lravgprs = rtaxso) if year==1995, r;

IV (2SLS) regression with robust standard errors       Number of obs =      48
                                                       F(  2,    45) =    8.19
                                                       Prob > F      =  0.0009
                                                       R-squared     =  0.4189
                                                       Root MSE      =  .18957

------------------------------------------------------------------------------
             |               Robust
     lpackpc |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    lravgprs |  -1.143375   .3723025    -3.07   0.004    -1.893231   -.3935191
     lperinc |    .214515   .3117467     0.69   0.495     -.413375     .842405
       _cons |   9.430658   1.259392     7.49   0.000     6.894112     11.9672
------------------------------------------------------------------------------
Instrumented:  lravgprs
Instruments:   lperinc rtaxso          STATA lists ALL the exogenous regressors
                                       as instruments – slightly different
                                       terminology than we have been using
------------------------------------------------------------------------------
```

- Running IV as a single command yields correct *SE*s
- Use , r for heteroskedasticity-robust *SE*s

Example: Cigarette demand, two instruments

```
             Y      W        X        Z1      Z2
. ivreg lpackpc lperinc (lravgprs = rtaxso rtax) if year==1995, r;

IV (2SLS) regression with robust standard errors       Number of obs =      48
                                                       F(  2,     45) =   16.17
                                                       Prob > F      =  0.0000
                                                       R-squared     =  0.4294
                                                       Root MSE      =  .18786

------------------------------------------------------------------------------
             |               Robust
     lpackpc |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    lravgprs |  -1.277424   .2496099    -5.12   0.000    -1.780164   -.7746837
     lperinc |   .2804045   .2538894     1.10   0.275     -.230955    .7917641
       _cons |   9.894955   .9592169    10.32   0.000     7.962993    11.82692
------------------------------------------------------------------------------
Instrumented:  lravgprs
Instruments:   lperinc rtaxso rtax   STATA lists ALL the exogenous regressors
                                     as "instruments" – slightly different
                                     terminology than we have been using
------------------------------------------------------------------------------
```

TSLS estimates, Z = sales tax (m = 1)

$$\widehat{\ln}(Q_i^{cigarettes}) = \underset{(1.26)}{9.43} - \underset{(0.37)}{1.14}\widehat{\ln}(P_i^{cigarettes}) + \underset{(0.31)}{0.21}\ln(Income_i)$$

TSLS estimates, Z = sales tax, cig-only tax (m = 2)

$$\widehat{\ln}(Q_i^{cigarettes}) = \underset{(0.96)}{9.89} - \underset{(0.25)}{1.28}\widehat{\ln}(P_i^{cigarettes}) + \underset{(0.25)}{0.28}\ln(Income_i)$$

- Smaller *SEs* for *m* = 2. Using 2 instruments gives more information – more “as-if random variation”.
- Low income elasticity (not a luxury good); income elasticity not statistically significantly different from 0
- Surprisingly high price elasticity

The General Instrument Validity Assumptions

$Y_i = \beta_0 + \beta_1 X_{1i} + \ldots + \beta_k X_{ki} + \beta_{k+1} W_{1i} + \ldots + \beta_{k+r} W_{ri} + u_i$

(1) ***Instrument exogeneity***: corr(Z_{1i},u_i) = 0,…, corr(Z_{mi},u_i) = 0

(2) ***Instrument relevance***: *General case, multiple X's*

Suppose the second stage regression could be run using the predicted values from the *population* first stage regression. Then: there is no perfect multicollinearity in this (infeasible) second stage regression.

- Multicollinearity interpretation…
- *Special case of one X*: the general assumption is equivalent to (a) at least one instrument must enter the population counterpart of the first stage regression, and (b) the *W*'s are not perfectly multicollinear.

The IV Regression Assumptions

$$Y_i = \beta_0 + \beta_1 X_{1i} + \ldots + \beta_k X_{ki} + \beta_{k+1} W_{1i} + \ldots + \beta_{k+r} W_{ri} + u_i$$

1. $E(u_i|W_{1i},\ldots,W_{ri}) = 0$
   - #1 says “the exogenous regressors are exogenous.”
2. $(Y_i,X_{1i},\ldots,X_{ki},W_{1i},\ldots,W_{ri},Z_{1i},\ldots,Z_{mi})$ are i.i.d.
   - #2 is not new
3. The *X*’s, *W*’s, *Z*’s, and *Y* have nonzero, finite 4th moments
   - #3 is not new
4. The instruments $(Z_{1i},\ldots,Z_{mi})$ are valid.
   - We have discussed this

- Under 1-4, TSLS and its *t*-statistic are normally distributed
- The critical requirement is that the instruments be valid…

Checking Instrument Validity

Recall the two requirements for valid instruments:

1. *Relevance* (special case of one X)

   At least one instrument must enter the population counterpart of the first stage regression.

2. *Exogeneity*

   All the instruments must be uncorrelated with the error term: corr(Z_{1i},u_i) = 0,…, corr(Z_{mi},u_i) = 0

What happens if one of these requirements isn't satisfied?

How can you check? What do you do?

If you have multiple instruments, which should you use?

Checking Assumption #1: Instrument Relevance

We will focus on a single included endogenous regressor:

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 W_{1i} + \ldots + \beta_{1+r} W_{ri} + u_i$$

First stage regression:

$$X_i = \pi_0 + \pi_1 Z_{1i} + \ldots + \pi_m Z_{mi} + \pi_{m+1} W_{1i} + \ldots + \pi_{m+k} W_{ki} + u_i$$

- The instruments are relevant if at least one of $\pi_1, \ldots, \pi_m$ are nonzero.
- The instruments are said to be ***weak*** if all the $\pi_1, \ldots, \pi_m$ are either zero or nearly zero.
- ***Weak instruments*** explain very little of the variation in X, beyond that explained by the W's

What are the consequences of weak instruments?

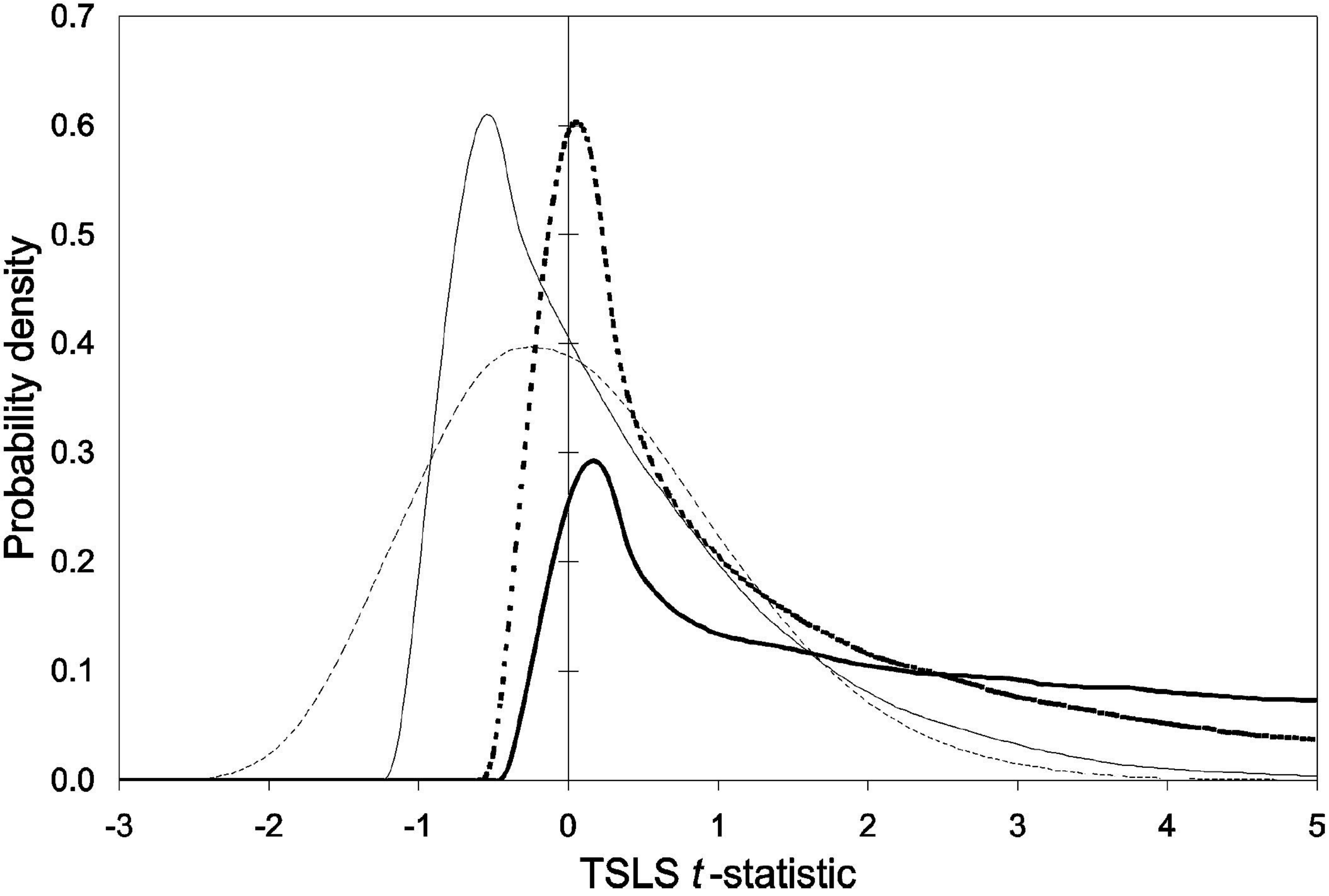
If instruments are weak, the sampling distribution of TSLS and its *t*-statistic are not (at all) normal, even with *n* large. Consider the simplest case:

$$Y_i = \beta_0 + \beta_1 X_i + u_i$$

$$X_i = \pi_0 + \pi_1 Z_i + u_i$$

- The IV estimator is $\hat{\beta}_1^{TSLS} = \dfrac{s_{YZ}}{s_{XZ}}$
- If cov(X,Z) is zero or small, then s_{XZ} will be small: With weak instruments, the denominator is nearly zero.
- If so, the sampling distribution of $\hat{\beta}_1^{TSLS}$ (and its *t*-statistic) is not well approximated by its large-*n* normal approximation…

An example: the sampling distribution of the TSLS *t*-statistic with weak instruments



Dark line = irrelevant instruments

Dashed light line = strong instruments

Why does our trusty normal approximation fail us?

$$\hat{\beta}_1^{TSLS} = \frac{s_{YZ}}{s_{XZ}}$$

- If cov(X,Z) is small, small changes in s_{XZ} (from one sample to the next) can induce big changes in $\hat{\beta}_1^{TSLS}$
- Suppose in one sample you calculate s_{XZ} = .00001...
- Thus the large-n normal approximation is a poor approximation to the sampling distribution of $\hat{\beta}_1^{TSLS}$
- A better approximation is that $\hat{\beta}_1^{TSLS}$ is distributed as the *ratio* of two correlated normal random variables (see SW App. 12.4)
- If instruments are weak, the usual methods of inference are unreliable – potentially very unreliable.

Measuring the strength of instruments in practice: The first-stage *F*-statistic

- The first stage regression (one X):

  Regress X on $Z_1,..,Z_m,W_1,\ldots,W_k$.
- Totally irrelevant instruments ⇔ *all* the coefficients on $Z_1,\ldots,Z_m$ are zero.
- The ***first-stage F-statistic*** tests the hypothesis that $Z_1,\ldots,Z_m$ do not enter the first stage regression.
- Weak instruments imply a small first stage F-statistic.

Checking for weak instruments with a single *X*

- Compute the first-stage *F*-statistic.

  Rule-of-thumb: If the first stage F-statistic is less than 10, then the set of instruments is weak.
- If so, the TSLS estimator will be biased, and statistical inferences (standard errors, hypothesis tests, confidence intervals) can be misleading.
- Note that simply rejecting the null hypothesis that the coefficients on the *Z*'s are zero isn't enough – you actually need substantial predictive content for the normal approximation to be a good one.
- There are more sophisticated things to do than just compare *F* to 10 but they are beyond this course.

What to do if you have weak instruments?

- Get better instruments (!)

- If you have many instruments, some are probably weaker than others and it’s a good idea to drop the weaker ones (dropping an irrelevant instrument will increase the first-stage *F*)

Estimation with weak instruments

- There are no consistent estimators if instruments are weak or irrelevant.
- However, some estimators have a distribution more centered around β_1 than does TSLS
- One such estimator is the limited information maximum likelihood estimator (LIML)
- The LIML estimator
  - can be derived as a maximum likelihood estimator

Checking Assumption #2: Instrument Exogeneity

- Instrument exogeneity: ***All*** the instruments are uncorrelated with the error term: corr(Z_{1i},u_i) = 0,…, corr(Z_{mi},u_i) = 0
- If the instruments are correlated with the error term, the first stage of TSLS doesn't successfully isolate a component of X that is uncorrelated with the error term, so $\hat{X}$ is correlated with u and TSLS is inconsistent.
- If there are more instruments than endogenous regressors, it is possible to test – *partially* – for instrument exogeneity.

Testing overidentifying restrictions

Consider the simplest case:

$$Y_i = \beta_0 + \beta_1 X_i + u_i,$$

- Suppose there are two valid instruments: Z_{1i}, Z_{2i}
- Then you could compute two separate TSLS estimates.
- Intuitively, if these 2 TSLS estimates are very different from each other, then something must be wrong: one or the other (or both) of the instruments must be invalid.
- The *J*-test of overidentifying restrictions makes this comparison in a statistically precise way.
- This can only be done if #*Z*'s > #*X*'s (overidentified).

Suppose #instruments = $m >$ # X's = k (overidentified)

$$Y_i = \beta_0 + \beta_1 X_{1i} + \ldots + \beta_k X_{ki} + \beta_{k+1} W_{1i} + \ldots + \beta_{k+r} W_{ri} + u_i$$

The *J*-test of overidentifying restrictions

The *J*-test is the Anderson-Rubin test, using the TSLS estimator instead of the hypothesized value $\beta_{1,0}$. The recipe:

1. First estimate the equation of interest using TSLS and all m instruments; compute the predicted values $\hat{Y}_i$, using the *actual* X's (not the $\hat{X}$'s used to estimate the second stage)
2. Compute the residuals $\hat{u}_i = Y_i - \hat{Y}_i$
3. Regress $\hat{u}_i$ against $Z_{1i},\ldots,Z_{mi}$, $W_{1i},\ldots,W_{ri}$
4. Compute the F-statistic testing the hypothesis that the coefficients on $Z_{1i},\ldots,Z_{mi}$ are all zero;
5. The ***J-statistic*** is $J = mF$

$J = mF$, where F = the F-statistic testing the coefficients on $Z_{1i},\ldots,Z_{mi}$ in a regression of the TSLS residuals against $Z_{1i},\ldots,Z_{mi}$, $W_{1i},\ldots,W_{ri}$.

Distribution of the *J*-statistic

- Under the null hypothesis that all the instruments are exogeneous, J has a chi-squared distribution with $m–k$ degrees of freedom
- If $m = k$, $J = 0$ (*does this make sense?*)
- If some instruments are exogenous and others are endogenous, the J statistic will be large, and the null hypothesis that all instruments are exogenous will be rejected.

Checking Instrument Validity: Summary

The two requirements for valid instruments:

1. ***Relevance*** (special case of one X)

- At least one instrument must enter the population counterpart of the first stage regression.
- If instruments are weak, then the TSLS estimator is biased and the and *t*-statistic has a non-normal distribution
- To check for weak instruments with a single included endogenous regressor, check the first-stage *F*
  - If F>10, instruments are strong – use TSLS
  - If F<10, weak instruments – take some action

2. *Exogeneity*

- ***All*** the instruments must be uncorrelated with the error term: corr(Z_{1i},u_i) = 0,…, corr(Z_{mi},u_i) = 0
- We can partially test for exogeneity: if m>1, we can test the hypothesis that all are exogenous, against the alternative that as many as m–1 are endogenous (correlated with u)
- The test is the *J*-test, constructed using the TSLS residuals.